CLAIMS

What is claimed is:

1. A method for assessing similarity between two data objects, comprising the steps of:
    a. receiving two data objects of type X;
    b. deriving respective variables for each of said two data objects;
    c. comparing the respective variables of said two data objects to derive an X,X comparison; and
    d. running said X,X comparison through a predictive model to calculate a similarity score for said two data objects.

2. The method of claim 1, wherein said predictive model comprises a neural network.

3. The method of claim 1, wherein said predictive model comprises a regression model.

4. The method of claim 1, wherein the two data objects include documents.

5. The method of claim 1, wherein the two data objects include one of resumes and job descriptions.

6. The method of claim 5, wherein the respective derived variables include one or more of following:
    i. reduced representation of the words in a resume;
    ii. reduced representation of the words in the education section of a resume;
    iii. reduced representation of each job description in a resume;
    iv. years of experience;
    v. standardized variables;
    vi. such as job titles;
    vii. industry SIC codes; and
    viii. degree names.

7. The method of claim 1, further comprising the steps of:

e. repeating steps a) through d) for a plurality of data objects of type X; and

f. clustering the plurality of data objects according to the calculated similarity scores.

8. The method of claim 1, further comprising the steps of:

e. repeating steps a) through d) for a plurality of data objects of type X in a database; and

f. organizing the database of objects of type X based on the calculated similarity scores.

9. The method of claim 1, further comprising the steps of:

e. repeating steps a) through d) for a plurality of data objects of type X; and

f. deriving from the calculated similarity scores one of a measure of supply of data objects of type X and a measure of demand for a particular one of the plurality of data objects of type X.

10. The method of claim 1, wherein the step of comparing the respective variables further includes the steps of:

i) constructing a vector for each of the two data objects from the derived respective variables; and

ii) calculating the cosine of the angle between the vectors.

11. The method of claim 1, wherein the step of comparing the respective variables further includes the steps of:

i) constructing a vector for each of the two data objects from the derived respective variables; and

ii) calculating the dot product of the vectors.

12. A method for assessing similarity between two data objects, comprising the steps of:

a. training a first predictive model with a first set of data objects of type X and matched data objects of type Y;

b. using said first predictive model to assess compatibility between each of a plurality of X,Y pairs, wherein for each X,Y pair, each X is a member of a second set of data objects of type X and each Y is a member of a second set of data objects of type Y;

c. assigning an X,Y compatibility score to each X,Y pair;

d. comparing the X,Y compatibility scores of each member of the second set of data objects of type X with each other member of the second set of data objects of type X;

e. pairing each member of the second set of data objects of type X with selected other members of the second set of data objects of type X having similar X,Y compatibility scores to identify a first plurality of X,X pairs, said first plurality of X,X pairs being matched pairs for training a second predictive model;

f. selecting other ones of the second set of data objects of type X that do not have as similar compatibility scores as the matched pairs to identify a second plurality of X,X pairs, said second plurality of X,X pairs being distracters for training said second predictive model;

g. deriving a respective set of variables from each member of the second set of data objects of type X;

h. comparing the respective set of variables derived from each X,X matched pair and from each X,X distracter pair to determine a set of X,X comparisons;

i. training a second predictive model with said set of X,X comparisons;

j. receiving two data objects of type X that are not in either the first training dataset or second training dataset;

k. deriving respective variables from each of said two data objects of type X;

l. comparing the respective variables derived from each of said two data objects of type X to determine a production X,X comparison; and

m. running said production X,X comparison through said second predictive model to calculate a similarity score for said two data objects of type X.

13. The method of claim 12, wherein either of said first and second predictive models comprise a respective neural network.

14. The method of claim 12, wherein either of said first and second predictive models comprise a respective regression model.

15. The method of claim 12, wherein the two data objects of type X include documents.

16. The method of claim 12, wherein the two data objects of type X include one of resumes and job descriptions.

17. The method of claim 16, wherein the respective derived variables and the respective sets of derived variables include one or more of following:

i. reduced representation of the words in a resume;

ii. reduced representation of the words in the education section of a resume;

iii. reduced representation of each job description in a resume;

iv. years of experience;

v. standardized variables;

vi. such as job titles;

vii. industry SIC codes; and

viii. degree names.

18. The method of claim 12, further comprising the steps of:

n. repeating steps j) through m) for a plurality of production data objects of type X not in the first and second training datasets; and

o. clustering the plurality of production data objects according to the calculated similarity scores.

19. The method of claim 12, further comprising the steps of:

n. repeating steps j) through m) for a plurality of production data objects of type X in a database; and

o. organizing the database of production data objects of type X based on the calculated similarity scores.

20. The method of claim 12, further comprising the steps of:

n. repeating steps j) through m) for a plurality of production data objects of type X; and

o. deriving from the calculated similarity scores one of a measure of supply of data objects of type X and a measure of demand for a particular one of the plurality of the production data objects of type X.

21. The method of claim 12, wherein the step of comparing the respective variables further includes the steps of:

i) constructing a vector for each of the two data objects from the derived respective variables; and

ii) calculating the cosine of the angle between the vectors.

22. The method of claim 12, wherein the step of comparing the respective variables further includes the steps of:

i) constructing a vector for each of the two data objects from the derived respective variables; and

ii) calculating the dot product of the vectors.

23. A method for assessing similarity between two data objects, comprising the steps of:

a. training a predictive model with a first set of data objects of type X and matched data objects of type Y, such that data objects of type X are treated the same as data objects of type Y with respect one or more common attributes;;

b. receiving two data objects of type X that are not in the first set of data objects; and

c. running said received two data objects of type X through said predictive model as though one were a data object of type X and the other were a data object of type Y to calculate a similarity score for said received two data objects of type X.

24. The method of claim 23, wherein the one or more common attributes includes considering one of the data objects as a bag of words.

25. The method of claim 23, wherein said predictive model comprises a neural network.

26. The method of claim 23, wherein said predictive model comprises a regression model.

27. The method of claim 23, wherein the received two data objects include documents.

28. The method of claim 23, wherein the received two data objects include one of resumes and job descriptions.

29. The method of claim 23, further comprising the steps of:

d. repeating steps b) and c) for a plurality of data objects of type X; and

e. clustering the plurality of data objects according to the calculated similarity scores.

30. The method of claim 23, further comprising the steps of:

d. repeating steps b) and c) for a plurality of data objects of type X in a database; and

e. organizing the database of objects of type X based on the calculated similarity scores.

31. The method of claim 23, further comprising the steps of:

d. repeating steps b) through c) for a plurality of data objects of type X; and

e. deriving from the calculated similarity scores one of a measure of supply of data objects of type X and a measure of demand for a particular one of the plurality of data objects of type X.

32. A system for assessing similarity between two data objects, comprising:

a. means for assessing compatibility between data objects of type X and data objects of type Y;

b. means for assessing similarity of a pair of data objects of type X, wherein the pair includes a first data object of type X and a second data object of type X and the assessed similarity of the pair is based on the compatibility of each of the first and second data objects with objects of type Y;

c. means for deriving respective variables from each of the first and second data objects of type X;

d. means for comparing the respective variables to calculate a comparison of the first and second data objects; and

e. a predictive model trained with the comparison of the respective variables derived from the first and second data objects of type X.

33. The system of claim 32, wherein said predictive model comprises a neural network.

34. The system of claim 32, wherein said predictive model comprises a regression model.

35. The system of claim 32, wherein the first and second data objects include documents.

36. The system of claim 32, wherein the first and second data objects include one of resumes and job descriptions.